Website:https://worldsiencepub.com/index.php/Imc Volume:5,Issue 02,September-2026

WORD ERROR RATE, CHARACTER ERROR RATE, AND PHONEME ERROR RATE IN ASR: WHAT EACH METRIC MEASURES AND WHERE THEY DIVERGE

A Brief Comparative Study

Utkirbek Azimov
Independent researcher
ORCID: 0000-0002-8088-324X

September 2026
DOI:https://doi.org/10.5281/zenodo.23027561

Abstract: Word error rate (WER), character error rate (CER), and phoneme error rate (PER) measure different aspects of automatic speech recognition (ASR). Controlled Uzbek examples illustrate how orthographic variation can raise WER without changing PER, while loanwords and allophony can yield higher naive PER than CER. An Acceptance layer distinguishes permissible pronunciation variants from errors. These illustrative findings support complementary evaluation, with explicit documentation of G2P and acceptance criteria rather than population-level conclusions.

Keywords: automatic speech recognition; word error rate; character error rate; phoneme error rate; grapheme-to-phoneme; Uzbek; agglutinative languages

1. Introduction

WER, CER, and PER compare reference and hypothesis sequences at word, character, and phoneme levels [1,2]. Their disagreement reveals different discrepancies rather than interchangeable measures of quality. Uzbek provides a useful setting because agglutination, apostrophe variation, and loanword phonetics complicate evaluation. This study explains these differences using controlled, fixed test sets.

2. The Three Metrics

2.1. Word Error Rate (WER)

WER divides word substitutions (S), deletions (D), and insertions (I) by the reference word count (N_w):

$$WER = (S + D + I) / N_w.$$

Each incorrect, omitted, or additional word contributes an error. WER measures transcription accuracy but depends on tokenization, punctuation, case, and orthographic conventions [1]. In Uzbek, attaching versus separating suffixes can change WER without changing meaning. For reproducible comparisons, these conventions must be fixed rather than assumed equivalent across corpora.

2.2. Character Error Rate (CER)

CER applies the same calculation to characters [1]. It is less sensitive to word boundaries: separating a suffix may create word-level errors but only a small character discrepancy. However, character distance does not directly measure phonetic or semantic significance. It therefore complements WER when word boundaries are ambiguous, but a small spelling change need not imply a minor pronunciation error.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

2.3. Phoneme Error Rate (PER)

Here, reference and hypothesis texts pass through a grapheme-to-phoneme (G2P) converter. PER divides their phoneme edit distance by the reference phoneme count, measuring differences in implied pronunciation [2]. It informs pronunciation, accent, loanword, and morphological analysis, but depends on G2P accuracy and phoneme conventions. Identical recognition errors can receive different PER values under different pipelines. Its additional dependence on the converter makes PER particularly fragile and demands fuller pipeline documentation than word- or character-level scoring.

3. Where the Metrics Diverge

3.1. Orthographic variation and WER

In a controlled Uzbek test set of 4,099 samples, imposed orthographic variation increased WER by approximately 8.3 percentage points while PER remained zero; CER fell between these behaviors. Word alignment registered surface differences without changed phonetic output. Consequently, WER comparisons require consistent orthography, tokenization, and apostrophe handling. PER can remove such discrepancies but also hides the surface errors WER captures. Even identical reported WER values, such as 15%, need not be comparable when different apostrophe conventions contribute different amounts of error.

3.2. Phonetic divergence and permissible variants

Among 30 selected Uzbek examples involving Russian loanwords and allophony, CER was 12.42, naive PER 21.87, and PER after a verified Acceptance layer 0.00. Character-close forms showed apparent phonemic discrepancies involving vowel quality, allophone mapping, or pronunciation sequences. Naive PER highlighted phonologically sensitive material that CER made less conspicuous.

However, apparent phoneme mismatches can be permissible variants rather than genuine errors. The reduction to zero after acceptance filtering demonstrates how strongly variant handling affects this selected slice. These results are illustrative, not general population-level estimates. Higher naive PER should therefore not automatically be interpreted as more genuine pronunciation errors: the Acceptance layer separates permissible variation from mismatches that count as errors.

4. Implications for Uzbek

Different Unicode apostrophe-like characters associated with o‘ and g‘ can produce corpus-dependent discrepancies [3]. PER reduces this sensitivity only if G2P consistently canonicalizes variants; otherwise, it inherits the inconsistency. Agglutination likewise makes WER sensitive to suffix boundaries and CER to spelling length. PER can reduce orthographic effects only when G2P correctly handles suffix phonology across contexts.

Loanwords and allophony can produce pronunciation-level differences disproportionate to character changes. Interpreting them requires a fixed phoneme inventory and pronunciation-acceptance criteria defined in advance.

5. Practical Recommendations

Report all three metrics on the same test set and references, with consistent, documented tokenization and normalization.

Specify the G2P system and version, phoneme inventory, stress conventions, digraph and affricate handling, apostrophe canonicalization, and representations of o‘ and g‘. Undocumented or different pipelines prevent direct PER comparison.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

Report naive and accepted PER using predefined permissible variants. Retain WER and CER as central transcription metrics; use PER diagnostically for text-consuming tasks and centrally where phoneme-level recognition, synthesis, or pronunciation is the focus.

Use synthetic controls to isolate orthographic artifacts from linguistic errors, including tokenization and punctuation effects. PER's standardization difficulties require documentation, not abandonment.

6. Conclusion

WER measures word-level surface accuracy, CER character-level accuracy, and PER distance between implied pronunciations. Orthography can raise WER without affecting PER, while phonemic differences and permissible variants complicate CER–PER comparisons. No single metric is sufficient. With fixed G2P and acceptance criteria, PER complements rather than replaces WER and CER. Their divergence reveals information about systems and data, not noise to average away.

References

1. National Institute of Standards and Technology (NIST). (2022). OpenASR21 Challenge Results. NIST. https://www.nist.gov/itl/tted/mltg/openasr21-challenge-results
2. Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, 50(5), 434–451. https://doi.org/10.1016/j.specom.2008.01.002
3. Unicode Consortium. (2024). Unicode Common Locale Data Repository (CLDR), Version 46: Uzbek locale data and collation. https://www.unicode.org/cldr/charts/46/collation/uz.html
4. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics–Doklady, 10(8), 707–710. Translated from Doklady Akademii Nauk SSSR, 163(4), 845–848 (1965).

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.